1. **Initial data exploration commands**

**\* Load the built-in dataset 'auto'**

**sysuse auto, clear**

**\* View the first ten rows**

**list in 1/10**

**\* Browse the data**

**browse**

**\* Describe the dataset**

**describe**

**\* Summarize the data**

**summarize**

**summarize, detail**

**\* Check for missing values**

**misstable summarize**

**\* Details of each variable**

**codebook**

**\* Frequency distribution for the 'rep78' variable (analogous to education)**

**tabulate rep78**

**tabulate rep78, missing**

**\* Histogram for the 'length' variable (analogous to income)**

**histogram length, normal**

**\* Dot plot for the 'length' variable**

**dotplot length**

**\* Density plot of the 'length' variable**

**kdensity length**

**\* Box plot for the 'price' variable**

**graph box price**

**\* Scatter plot for 'price' vs. 'mpg'**

**scatter price mpg**

**graph twoway (scatter price mpg) (lfit price mpg)**

**\* Correlation matrix**

**corr mpg price weight**

1. **Outlier identification and handling**

**\* Load the built-in dataset 'auto'**

**sysuse auto, clear**

**\* Box plot for 'price'**

**graph box price**

**\* Scatter plot for 'price'**

**gen id = \_n**

**scatter price id**

**\* Summarize 'price' to get basic statistics**

**summarize price, detail**

**\* Calculate IQR**

**gen iqr\_price = r(p75) - r(p25)**

**\* Identify outliers**

**\* data < (Q1 - 1.5 \* IQR) OR data > (Q1 + 1.5 \* IQR)**

**gen outlier = (price < (r(p25) - 1.5 \* iqr\_price)) | (price > (r(p75) + 1.5 \* iqr\_price))**

**\* List outliers**

**br make price if outlier**

**\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\***

**\* Handling outliers**

**\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\***

**\*(1) Setting to lower and upper range or IQR outlier formula**

**gen price\_capped = price**

**replace price\_capped = (r(p25) - 1.5 \* iqr\_price) if price < (r(p25) - 1.5 \* iqr\_price)**

**replace price\_capped = (r(p75) + 1.5 \* iqr\_price) if price > (r(p75) + 1.5 \* iqr\_price)**

**\* Verify the changes**

**graph box price\_capped**

**\*(2) Dropping outlier observations**

**drop if outlier == 1**

1. **Duplicates identification and handling**

**\* Load the built-in dataset 'auto'**

**sysuse auto, clear**

**\* Identify duplicate observations**

**duplicates report mpg trunk**

**\* List duplicate observations**

**duplicates list mpg trunk**

**\* Tag duplicates and create a variable 'dup\_tag'**

**duplicates tag mpg trunk, generate(dup\_tag) //dup\_tag show how many extra occurence is there.**

**\* browse observations with duplicates**

**br mpg trunk dup\_tag if dup\_tag**

**\* Drop duplicate observations**

**duplicates drop mpg trunk, force**

**\* Verify that duplicates are removed**

**duplicates report mpg trunk**

1. **Recoding and dummy variable generation**

**clear**

**set** seed 12345

**set** obs 50

**gen** id = \_n

**gen** income = round(runiform() \* 10000)

\*Recode income variable into categories

**recode** income (0/3000 = 1 "Low") (3001/7000 = 2 "Medium") (7001/max = 3 "High"), **generate**(income\_cat)

\*Generate dummy variable based on categorical variable

**tabulate** income\_cat, **gen**(inc)